United States
Department
of Agriculture

Forest Service

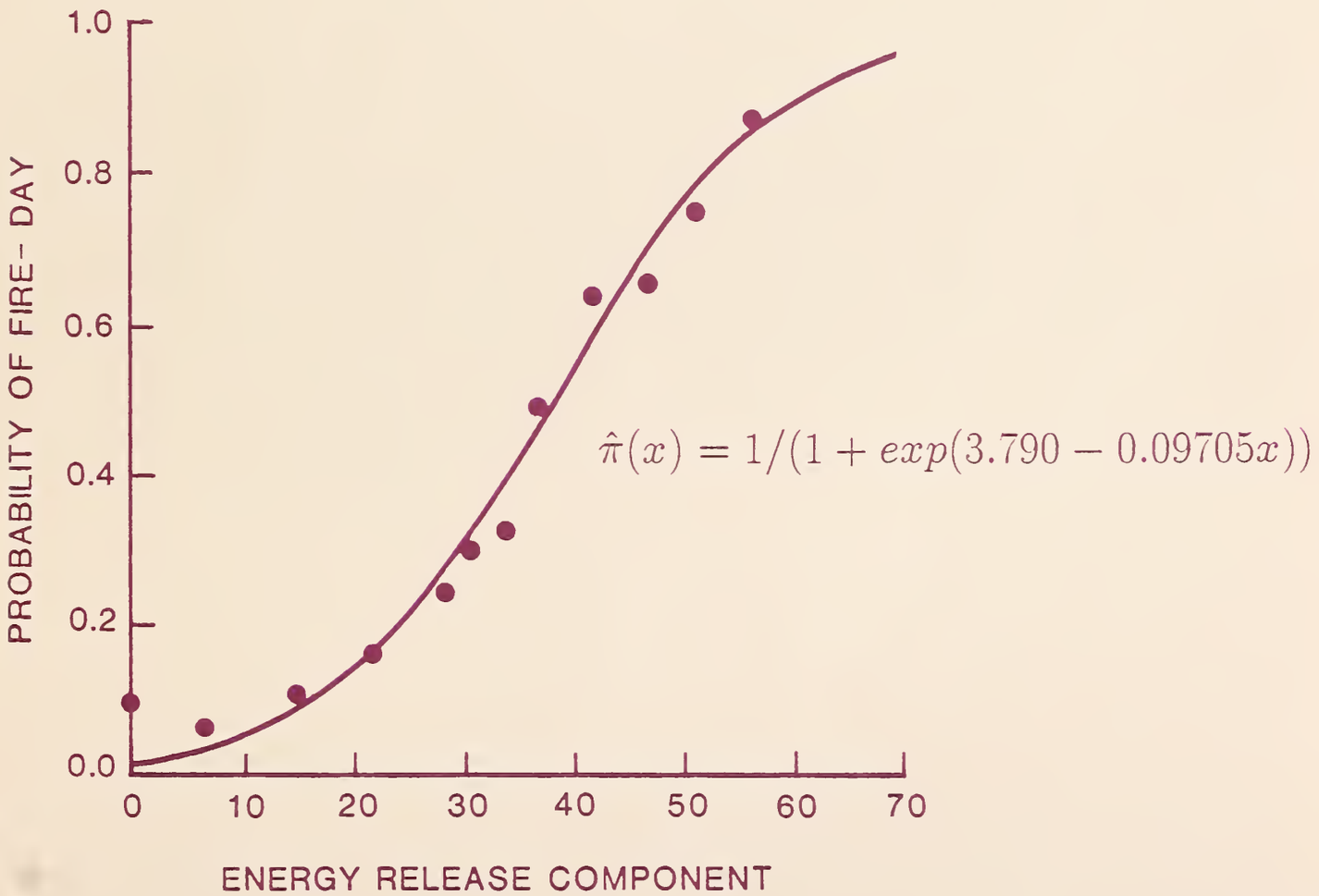
**Intermountain
Research Station**

General Technical
Report INT-286

May 1992

Constructing and Testing Logistic Regression Models for Binary Data: Applications to the National Fire Danger Rating System

Don O. Loftsgaarden
Patricia L. Andrews



THE AUTHORS

DON O. LOFTSGAARDEN received his bachelor's, master's, and Ph.D. degrees in mathematics (statistics emphasis) from Montana State University, the latter in 1964. He is currently Professor of Mathematics in the Department of Mathematical Sciences at the University of Montana. He teaches statistics courses at all levels and does statistical consulting for faculty and graduate students in many departments on campus. His interests are in the application of statistics. He has done statistical consulting for several State agencies, the Confederated Kootenai and Salish Tribes, the Conference Board of Mathematical Sciences, the American Mathematical Society, and others.

PATRICIA L. ANDREWS is a mathematician stationed at the Intermountain Fire Sciences Laboratory in Missoula, MT. She received her B.A. in mathematics and chemistry from Eastern Montana College, Billings, in 1970, and her M.A. in mathematics and computer science in 1973 from the University of Montana, Missoula. She has been a member of the Fire Behavior research work unit since 1973. Her work has centered on putting mathematical models into a form that can be easily used and understood by fire managers. She is now team leader of the systems development and applications team.

This research was supported in part by funds provided by the Intermountain Research Station, Forest Service, U.S. Department of Agriculture (Agreement No. INT-87228-COA).

CONTENTS

Intermountain Research Station
324 25th Street
Ogden, UT 84401

Constructing and Testing Logistic Regression Models for Binary Data: Applications to the National Fire Danger Rating System

Don O. Loftsgaarden
Patricia L. Andrews

INTRODUCTION

Logistic regression proved to be useful in examining the relationship between National Fire Danger Rating System (NFDRS) indexes and historical fire occurrence data. Specific results of the study will be reported elsewhere. This report presents selected examples to illustrate our statistical methodology, which we believe will be helpful to potential users of logistic regression. We assume that the reader has some basic knowledge of statistical methods.

Many applications of logistic regression are classical dose/response studies. An experiment is designed so that several dosages for the explanatory variable x are used. The response is then recorded as either a 0 or 1 (that is, dead or alive, cured or not cured, etc.) Our applications differ from this scenario in several important aspects:

- Because the explanatory variable is calculated from daily weather observations, we have no control over the "dosage."
- An NFDRS index can range from 0 to as high as 200, so it is necessary to group the data into categories for tests of fit.
- One part of our study involved comparing 100 logistic regression models generated from the same set of weather data (20 fuel models x 5 indexes). This problem requires quick and easy methods for doing tests of fit.

This paper is not an exhaustive treatise on all aspects of logistic regression. The emphasis is on tests of fit. Techniques for both single-explanatory variable models and multi-explanatory variable models are given. Various grouping methods and how logistic regression programs handle grouping is one of the major topics of this paper. Tests commonly used in logistic regression computer procedures are discussed, including an explanation of when they are not applicable. Constructing, evaluating, and comparing a large number of logistic models necessitated discarding some techniques that have been used successfully elsewhere (Landwehr and others 1984; Pregibón 1981). Logistic regression is currently an active area of research in statistics, and continued improvements in methodology can be expected over the next several years (for example, Hosmer and Lemeshow 1989).

NFDRS/FIRE OCCURRENCE STUDY

The National Fire Danger Rating System (NFDRS) was designed to indicate the level of fire danger for each day. It is based on a daily weather reading and site conditions defined for a broad area (Deeming and others 1977). Although NFDRS indexes are used in making fire management decisions, very little attention has focused on the relationship between NFDRS indexes and quantitative measures of the fire situation (Andrews 1987). The need to examine this relationship was pointed out in the Final Report on Fire Management Policy, May 5, 1989, written by a task force commissioned as a result of the severe fire season of 1988, some of the most notable fires being in Yellowstone National Park. The following is found in the report: "Validation of the relationship between current fire management information system components (i.e., drought index, energy release component, 1,000-hour fuel moisture, etc.) and actual fire occurrence, severity and size is needed."

In this study we examined the relationship between NFDRS indexes and fire occurrence. A day was classified as a fire-day if one or more fires were reported on that day. Otherwise it was a no-fire-day. With an NFDRS index as the explanatory variable, we have a situation that is ideal for logistic regression. Martell and others (1987) have also used logistic regression in Canada for problems similar to those discussed here.

In addition to examining the relationship between indexes and fire occurrence, we had several other specific applications to which we hoped to apply the methods:

1. Test the performance of NFDRS and define what NFDRS can do. The NFDRS was not designed to predict the behavior of individual fires. It gives an overall rating of the daily fire potential, a concept that is not easy to define and measure. The logistic regression models described in this paper give probability of a fire-day for a specific index value. These probability curves can be used to help a fire manager calibrate an index for a specific area and to set decision points.

2. Develop a method for choosing the appropriate fuel model and index for a particular location. NFDRS offers the fire manager a choice of 20 fuel models to define the area of concern. The choice is most often based on a description in terms of vegetation and on experience with what seems to work best. In addition, there are five components and indexes from which to choose. The fire manager needs an objective means of choosing the "best" fuel model and index.

3. Evaluate proposed improvements to NFDRS. In response to deficiencies pointed out in NFDRS, changes have been proposed to the system (Burgan 1988). We wanted a means to evaluate whether the 1988 NFDRS was an improvement over the 1978 NFDRS.

Our work with logistic regression models has shown that these models can be utilized in all cases we envisioned above. We are optimistic that logistic regression models will be successfully applied to a wide variety of questions related to NFDRS.

LOGISTIC REGRESSION OVERVIEW

Logistic regression is a popular statistical tool. The following are recent applications of logistic regression:

- An insect will live or die after a certain dose of spray (Stukel 1988).
- A tree will live or die after a fire (Ryan and Reinhardt 1988).
- A lightning strike will either start a fire or not (Latham and Schlieter 1989).
- At a given moisture content, a smoldering fire will continue to burn or go out (Hartford 1989).

Our goal is to present the basic techniques at a modest mathematical level. Because logistic regression is not as widely used and understood as ordinary linear regression, we will draw analogies and contrasts between the two. Readers should be familiar with ordinary linear regression, the coefficient of determination R^2, and related tests and inferences.

Simple linear regression uses data of the form (x,y) where x is an explanatory (independent) variable and y is a response (dependent) variable that takes values on a continuous scale. Logistic regression uses data of the same form except y is a binary variable having only two values, 0 or 1. In either case the explanatory variable x may be one variable or a vector of several variables.

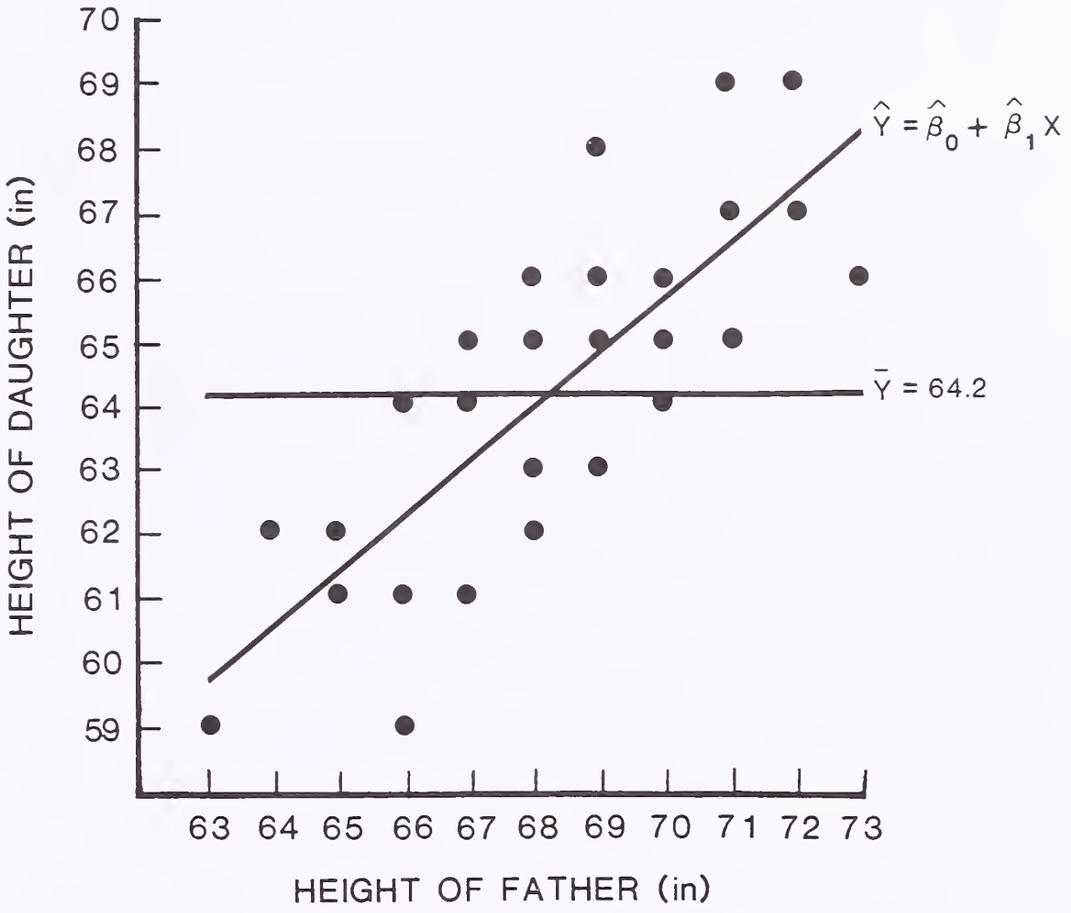


Figure 1—Linear regression example where the explanatory variable is father's height and the response variable is daughter's height.

We will briefly review ordinary linear regression in order to draw analogies between ordinary and logistic regression. An example of linear regression is shown in figure 1. The explanatory variable is father's height and the response

variable is daughter's height (full grown). The least-squares line is in the plot as well as a horizontal line at $\bar{Y}$. Here $R^2 = 0.63$. ($R^2 = 1$ if all observations fall on the line). Based on this model, if a father is 70 inches tall, we predict his daughter to be 65.7 inches tall.

An example of a logistic regression curve is shown in figure 2 (from Stukel 1989). The explanatory variable is dosage of insecticide, the response variable is insect death with values 0 (alive) or 1 (dead). The complete set of data and the logistic regression model are given in example 1 later in this paper. Each y value is either 0 or 1, and therefore none of the data pairs (x,y) fall on the curve. The 481 actual (x,y) pairs are not plotted in figure 2. There are approximately 60 observations at each of eight different dosages of insecticide. For an insecticide dosage of 1.8, the logistic model predicts that 0.72 of all insects that are sprayed will die.

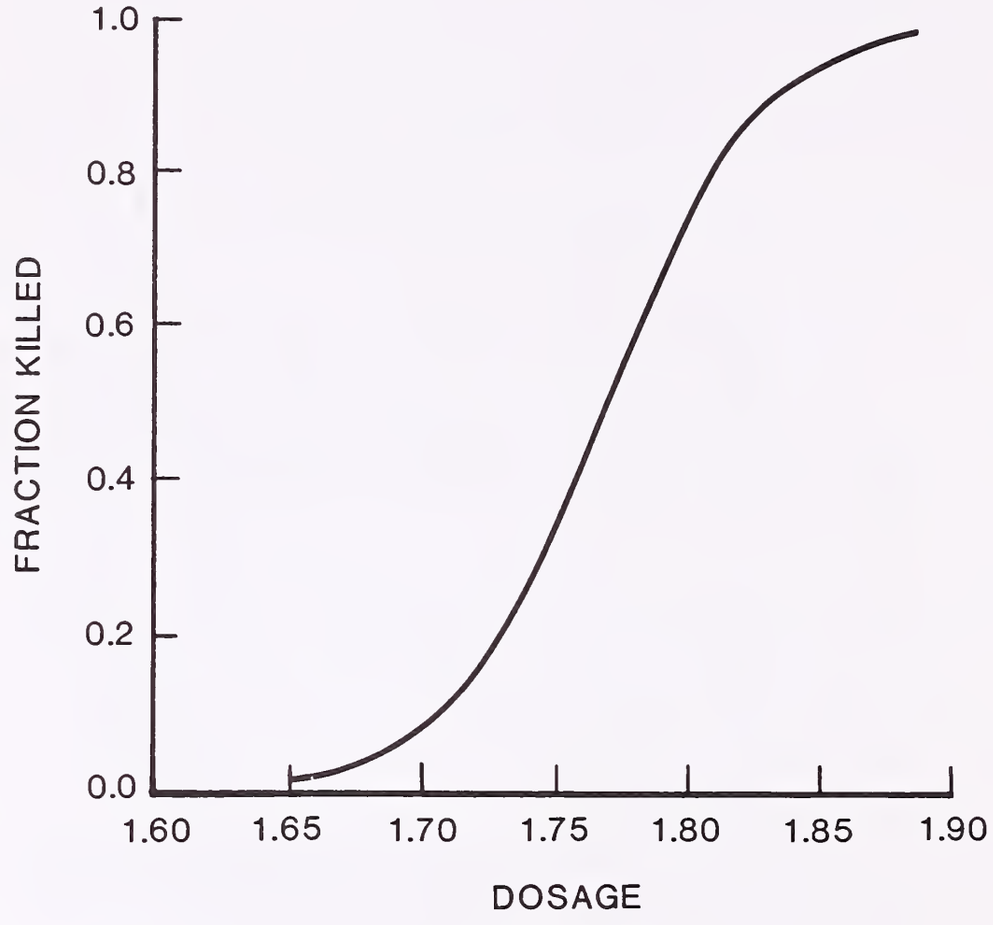


Figure 2—Logistic regression curve where the explanatory variable is dosage of insecticide and the response variable is insect death with values 0 (alive) or 1 (dead).

SIMPLE LINEAR REGRESSION REVIEW

Before discussing logistic regression and drawing analogies between the two, it will be useful to review a few facts about simple linear regression.

The simple linear regression model is

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, 2, ..., n$$

where β_0, β_1 are unknown parameters, x_i's are known constants, e_i's are independent random variables, and $E(e_i) = 0$, $\text{Var}(e_i) = \sigma^2$. (The e_i's are often assumed to be normally distributed.) The data are a random sample of n independent pairs $(x_1, y_1), ..., (x_n, y_n)$.

Consequences of the assumptions for a simple linear regression model are:

i) $E(Y \mid x) = \beta_0 + \beta_1 x$,
ii) $\text{Var}(Y \mid x) = \text{Var}(Y) = \sigma^2$ (a constant that does not depend on x), and
iii) $Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$ (if normal assumption for e's is made).

Least squares and maximum likelihood estimators for β_0, β_1 are found by minimizing

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2, \tag{1}$$

and are denoted by $\hat{\beta}_0, \hat{\beta}_1$. Also, predicted values are denoted as

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, 2, ..., n.$$

It is helpful to think of the following two models and residual variation of the Y_i's about each of these models:

Model 1:
$\hat{Y} = \bar{Y}$ (x is ignored)
$SSTOT = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$.

Model 2:
$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$.

Then the sum of squares explained by the regression model, model 2, is

$$SSR = SSTOT - SSE = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2.$$

The coefficient of determination is defined as $R^2 = SSR/SSTOT$. R^2 is the proportion of the total variation in Y_i's explained by model 2 (or explanatory variable x.)

This information is often summarized in an ANOVA (Analysis of Variance) table.

ANOVA

Source	df	*SS*	*P-value*
Regression	1	SSR	MSR
Residual	$n-2$	SSE	MSE
Total	$n-1$	$SSTOT$	

A fit of model 1 and model 2 to a set of data was shown earlier in figure 1.

Since $\hat{\beta}_0, \hat{\beta}_1$ were obtained by minimizing a sum of squared deviations, it is natural to measure variation about a model using squared deviations (squared distances from an observed Y_i to a predicted Y_i), as we did for the two models above.

A MODEL FOR BINARY DATA

In the introduction several examples were given where the response variable Y is a Bernoulli random variable. A Bernoulli random variable is either 0 or 1 where $P(Y = 1) = \pi$. $E(Y) = \pi \cdot 1 + (1 - \pi) \cdot 0 = \pi$ and $\text{Var}(Y) = \pi \cdot (1 - \pi)$. Often π depends on one or more explanatory variables. For one variable x, $\pi(x) = E(Y \mid x)$ is the probability that Y=1 as a function of x.

There are numerous problems if one tries to use the simple linear regression model for binary data. Some of these problems are

i) $E(Y \mid x) = \pi(x)$ is usually not linear,
ii) $\text{Var}(Y \mid x) = \pi(x) \cdot (1 - \pi(x))$ is not constant but depends on x,
iii) Y is not a normal random variable, and
iv) $\pi(x)$ is a probability and hence must have values between 0 and 1.

One approach that has been used is weighted least squares with simple linear regression (Haines and others 1983). The reason for weighted regression is to deal with the variances, which are not constant but depend on x. Nevertheless, most of the problems mentioned above still remain.

A logistic regression model has been used successfully in many recent applications. Here it is assumed that

$$\ell n(\pi(x)/(1 - \pi(x))) = \beta_0 + \beta_1 x.$$

Tests of fit discussed later in this paper are in one sense aimed at deciding whether models based on this assumption are realistic when compared to the actual data.

Solving the above expression for $\pi(x) = E(Y \mid x)$ we obtain:

$$\pi(x) = exp(\beta_0 + \beta_1 x)/(1 + exp(\beta_0 + \beta_1 x))$$

$$= 1/(1 + exp(-\beta_0 - \beta_1 x)).$$

Plotting $\pi(x)$ against x gives the logistic curve (see fig. 2). This curve is bounded below by 0 and above by 1.

In practice one has n independent observations (x_1, y_1),..., (x_n, y_n) to use in estimating β_0, β_1. For simple linear regression, if all of the data fall on the line, then $R^2 = 1$, but this can never happen for logistic regression. All of the Y values are either 0 or 1, so no observations fall on the curve (fig. 3).

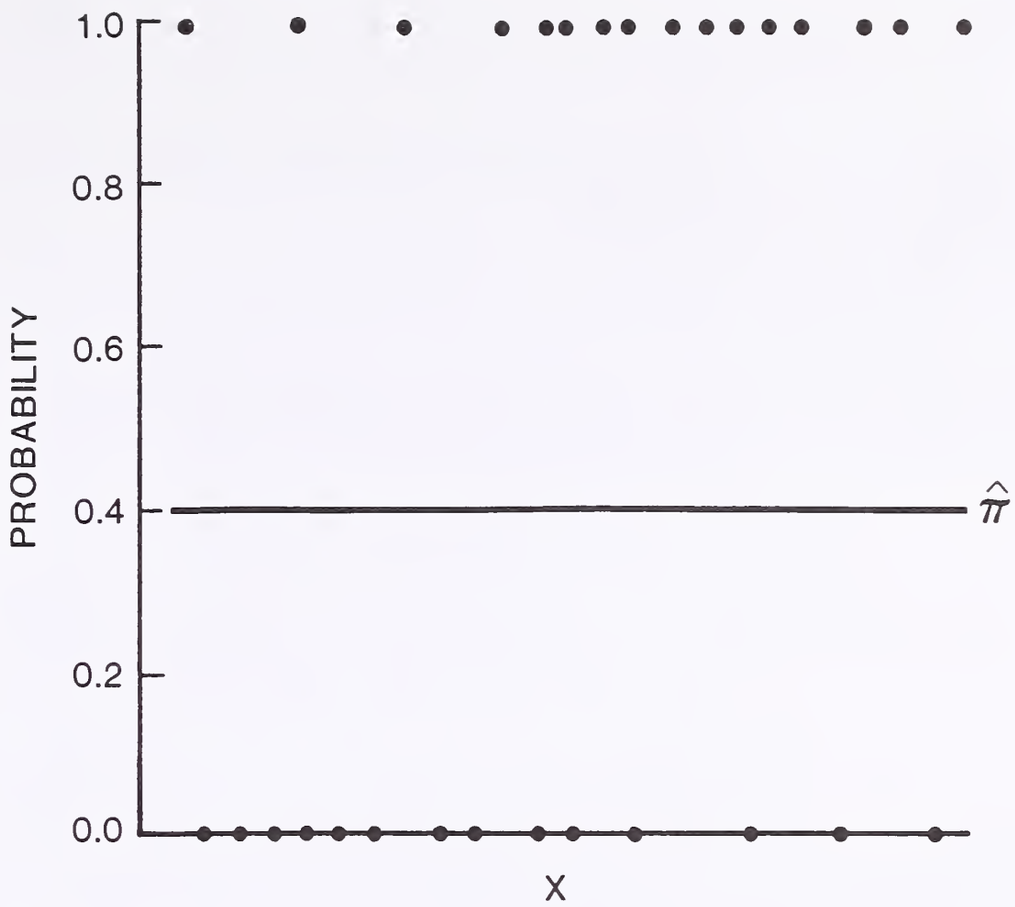


Figure 3—Plot of constant model, $\hat{\pi} = \bar{y}$ = overall fraction of 1's. Note that the y's are all 0's or 1's.

LOGISTIC REGRESSION—ONE EXPLANATORY VARIABLE

The Logistic Model

For logistic regression the response variable Y has probability distribution

y	0	1
$p(y)$	$1-\pi$	π

or

$$p(y) = \pi^y(1-\pi)^{1-y}, \quad y = 0, 1$$

with $E(Y) = \pi$.

It is assumed that π is a function of an explanatory variable x, that is, $\pi = \pi(x)$. Thus, $E(Y \mid x) = \pi(x)$. As shown earlier, when $\ell n(\pi(x)/(1-\pi(x)))$ is modeled as a linear function, $\beta_0 + \beta_1 x$, solving for $\pi(x)$ gives $\pi(x) = \mathrm{E(Y}|\, x) = 1/(1+exp(-\beta_0-\beta_1 x))$. The data consist of n independent pairs (x_1, y_1),..., (x_n, y_n). If $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimates for β_0 and β_1, then

$$\hat{\pi}(x) = 1/(1 + exp(-\hat{\beta}_0 - \hat{\beta}_1 x))$$

is the estimated logistic regression curve. For each x, $\hat{\pi}(x)$ is an estimate for the probability that $Y = 1$.

Maximum Likelihood Estimation

To understand how to extend the notion of sum of squares used for simple linear regression, one must examine the likelihood function for the logistic model given above.

$$p(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} \quad y_i = 0, 1.$$

where $\pi_i = \pi(x_i) = 1/(1 + exp(-\beta_0 - \beta_1 x_i))$, $i = 1, 2, \ldots, n$.
Let $\underline{\pi} = (\pi_1, ..., \pi_n)$ and $\underline{y} = (y_1, ..., y_n)$.
Then likelihood function, L, is

$$L = f(\underline{y} \mid \underline{\pi}) = \prod_{i=1}^{n} p(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}.$$

For finding maximum likelihood estimators it is easier to work with ℓnL,

$$\ell nL = \sum_{i=1}^{n}(y_i \ell n\pi_i + (1-y_i)\ell n(1-\pi_i)). \tag{2}$$

After replacing π_i by $1/(1 + exp(-\beta_0 - \beta_1 x_i))$ and simplification,

$$\ell nL = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} \ell n(1 + exp(\beta_0 + \beta_1 x_i)). \tag{3}$$

Also of interest is $-2\ell nL$. From (2),

$$-2\ell nL = \sum_{i=1}^{n} y_i(-2\ell n\pi_i) + \sum_{i=1}^{n}(1-y_i)(-2\ell n(1-\pi_i)). \tag{4}$$

Shortly we will see that $-2\ell nL$ will be used to give a statistic with known probability distribution.

Maximum likelihood estimates for β_0, β_1 are found by maximizing (3) as a function of β_0, β_1 or minimizing (4) (Cox 1989). If these estimates are $\hat{\beta}_0$ and $\hat{\beta}_1$, then $\hat{\pi}_i = \hat{\pi}(x_i) = 1/(1+exp(-\hat{\beta}_0 - \hat{\beta}_1 x_i))$. A measure of residual variation is found by substituting $\hat{\pi}_i$ for π_i in (4). This will be examined in detail in the following section.

Measuring Variability

If $0<\pi<1$ and y is 0 or 1, two measures of "closeness" or "distance between" π and y are now examined. S_1 and S_2 are defined as follows:

$$S_1(y,\pi) = \begin{cases} -2\ell n\pi & y = 1 \\ -2\ell n(1-\pi) & y = 0 \end{cases}$$

$$= -2[y\ell n\pi + (1-y)\ell n(1-\pi)] \qquad y = 0, 1$$

$$= [y(-2\ell n\pi) + (1-y)(-2\ell n(1-\pi))] \qquad y = 0, 1$$

$$S_2(y,\pi) = (y-\pi)^2 \qquad y = 0, 1.$$

Numerical examples are given in table 1. When y is near π, both S_1 and S_2 are small; and when y is far from π, both S_1 and S_2 are large. The motivation for S_1 comes from (4) (Efron 1978).

Table 1—Numerical values for S_1 and S_2

π	0.1	0.1	0.4	0.4	0.5	0.5	0.75	0.75	0.95	0.95
y	0	1	0	1	0	1	0	1	0	1
$S_1(y,\pi)$	0.211	4.605	1.022	1.833	1.386	1.386	2.773	0.575	5.991	0.103
$S_2(y,\pi)$	0.010	0.810	0.160	0.360	0.250	0.250	0.563	0.063	0.903	0.003

Consider $y_1, y_2, ..., y_n$ and corresponding $\pi_1, \pi_2, ..., \pi_n$. Recall that the y_i's are 0's and 1's and π_i's are probabilities. Two measures of variation of y_i's about π_i's are defined as follows:

$$\sum_{i=1}^{n} S_1(y_i, \pi_i) = -2 \sum_{i=1}^{n} [y_i \ell n \pi_i + (1 - y_i) \ell n (1 - \pi_i)], \quad (5)$$

$$\sum_{i=1}^{n} S_2(y_i, \pi_i) = \sum_{i=1}^{n} (y_i - \pi_i)^2. \quad (6)$$

Note that expressions (4) and (5) are the same; both are $-2\ell nL$. It is $-2\ell nL$ that is minimized to obtain $\hat{\beta}_0$ and $\hat{\beta}_1$, so it is the natural measure of variation to use. In the remainder of this paper (5) is used almost exclusively. It is analogous to the sum of squares for simple linear regression and it will be called a sum of squares.

There are two pieces in the sum on the righthand side of (5). The terms in the first piece are non-zero only for y_i's that are 1 and the terms in the second piece are non-zero only for y_i's that are 0. Formula (5) with $\hat{\pi}_i = 1/(1+exp(-\hat{\beta}_0 - \hat{\beta}_1 x))$ replacing π_i is analogous to $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$ for simple linear regression. Formula (5) can be used with $\hat{\pi}_i$'s from any model to obtain a measure of residual variation for that model. It is not limited to just the logistic regression model.

Closely related to S_2 and (6) is the idea of quadratic score for an observation and average score for a set of observations. In applications similar to ours, quadratic scores have been used inappropriately; such applications will be discussed farther on in this report.

Sums of Squares

Using (5) as a measure of variation of y_i's about a set of π_i's, it is possible to define "sums of squares" in a fashion similar to simple linear regression.

Starting with $\ell n(\pi/(1-\pi)) = \beta_0$ (no x in the model) and independent observations $y_1, y_2, ..., y_n$, the maximum likelihood estimator for β_0 is

$$\hat{\beta}_0 = \ell n(\overline{y}/(1-\overline{y})).$$

Equating this with $\ell n(\hat{\pi}/(1-\hat{\pi}))$ gives

$\hat{\pi} = \bar{y} = (\sum_{i=1}^{n} y_i)/n = (\#1\text{'s})/n$ =overall fraction of 1's.

The residual variation of $y_1, y_2, \ldots, y_n$ about the constant $\hat{\pi}$ is analogous to the total sum of squares for simple linear regression.

We define the total sum of squares as follows:

$$SSTOT = \sum_{i=1}^{n} S_1(y_i, \hat{\pi})$$
$$= -2[(\sum_{i=1}^{n} y_i)\ell n\hat{\pi} + (\sum_{i=1}^{n}(1 - y_i))\ell n(1 - \hat{\pi})]. \quad (7)$$

This simplifies to $SSTOT = -2[(\#1\text{'s})\ell n\hat{\pi} + (\#0\text{'s})\ell n(1-\hat{\pi})]$. For the remainder of this paper, $SSTOT$ will always be defined in this way.

For example, if $n = 200$, $\#1\text{'s} = 80$, $\#0\text{'s} = 120$, $\hat{\pi} = 80/200 = 0.4$, and $1 - \hat{\pi} = 0.6$, then $SSTOT = -2[80\ell n(0.4) + 120\ell n(0.6)] = 146.61 + 122.60 = 269.21$. This model is shown in figure 3. Notice that the y's are all 0's or 1's.

In general, we have a model $\hat{\pi}(x)$ that depends on x and is based on n independent pairs $(x_1, y_1), \ldots, (x_n, y_n)$. For each x_i we compute $\hat{\pi}_i = \hat{\pi}(x_i)$. A measure of residual variation of y_i's about $\hat{\pi}_i$'s is defined as follows:

$$SSE = \sum_{i=1}^{n} S_1(y_i, \hat{\pi}_i) = -2[\sum_{i=1}^{n} y_i \ell n\hat{\pi}_i + \sum_{i=1}^{n}(1 - y_i)\ell n(1 - \hat{\pi}_i)]. \quad (8)$$

Likewise, we define the sum of squares explained as $SSR = SSTOT - SSE$. SSR can be used to test Ho: $\beta_1 = 0$ against $H_a : \beta_1 \neq 0$ for a logistic model. When Ho is true, SSR has a chi-square distribution with 1 df and a large value of SSR leads us to reject Ho. All logistic regression packages we have used provide enough information to be able to carry out this test. We can use an ANOVA table to summarize this information.

ANOVA

Source	df	SS
Logistic model	1	SSR
Residual	$n - 2$	SSE
Total	$n - 1$	$SSTOT$

Classical Logistic Regression and ANOVA Table

Many of the original applications of logistic regression were to dose/response problems. For example, X might be the dosage of an insect spray and Y the response (Y=1 insect dead, Y=0 insect alive). Many observations are made at a small number of distinct dosages. Theoretically, as dosage increases the fraction of insects killed increases. An example from Stukel (1988) is discussed below.

In such problems, two models are considered:

Model 1: A logistic model with dosage as the explanatory variable.

Model 2: A best $\hat{\pi}$ model, which is defined as the fraction of 1's at each distinct value of x.

We denote these two models by $\hat{\pi}_1(x)$ and $\hat{\pi}_2(x)$, respectively. For model 1 we calculate SSE_1 and SSR_1 as described earlier. $SSTOT$ is the same for either model.

Model 2 is essentially the observed data. (For more details see McCullagh and Nelder 1989). Other models can be compared to model 2 as a test of how well the models fit the data. Model 2 is sometimes referred to as a saturated model because it has one parameter for each distinct value of the explanatory variable. Because model 2 essentially provides no reduction in the data, it is not really a model in the usual sense of the word.

It will be convenient to have some notation set up for model 2. Suppose that in the observations there are k distinct values of the explanatory variable, say $x_1,..., x_k$, that occur $m_1,..., m_k$ times, respectively, where $m_1 + ... + m_k =$ n and that the number of 1's at each x are $r_1, ..., r_k$, respectively. Then model 2 is:

$$\hat{\pi}_j = \frac{r_j}{m_j} \text{ at } x_j \text{ for } j = 1, 2, ..., k.$$

Substituting these values into (8) and simplifying gives SSE_2 for model 2 as

$$SSE_2 = -2[\textstyle\sum_{j=1}^{k} r_j \ell n \hat{\pi}_j + \sum_{j=1}^{k} (m_j - r_j) \ell n (1 - \hat{\pi}_j)].$$

Then $SSR_2 = SSTOT - SSE_2$. We note two further things here:

1. No other model has a smaller SSE than model 2. Since model 2 is essentially the observed data, this says that no model can agree better with the data than the data itself. This, of course, also means that SSR attains its largest value for model 2. It follows that $SSE_2 < SSE_1$ and $SSR_2 > SSR_1$.
2. If there are n observations and each observation has a distinct value of x (that is, $k = n$), then $SSE_2 = 0$ and $SSR_2 = SSTOT$.

We now consider a test of fit for model 1. Model 1 has two parameters, β_0 and β_1, that are estimated using all of the data. In contrast, model 2 has one parameter for each distinct value of x and each parameter is estimated using only the data for that particular value of x. If x has only a moderate number of distinct values and there are several observations at each value of x, then model 2 generally behaves as expected—it is monotonically increasing or decreasing as a function of x. In these cases, it is reasonable to compare model 1 with model 2 as a test of fit for model 1. On the other hand, if x has many distinct values (perhaps as many as n) and there are only a few observations at each value of x (perhaps only 1), then model 2 behaves erratically. In this situation, comparing model 1 to model 2 as a test of fit is uninformative and can be very misleading. In addition, the approximate probability distribution for the statistic to be used does not hold in this case (Hosmer and Lemeshow 1989; McCullagh and Nelder 1989).

The statistic to be used for the test of fit for model 1 is $SSR_2 - SSR_1$, which has an approximate chi-square distribution with $k - 2$ df. A small value of this statistic means that model 1 fits the data well and a large value indicates that model 1 does not fit the data well. Sometimes this statistic is called the deviance statistic. The preceding paragraph discusses when this statistic can

be used as a test of fit and when it is meaningless. Logistic regression software packages always seem to spit out this statistic so the user must be aware of when it is useful as a test-of-fit statistic and when it is not. Later we will discuss other tests of fit that can be used when this test is meaningless.

We consider one last item here. For ordinary regression, the coefficient of determination, R^2, is defined as $SSR/SSTOT$. For logistic regression we modify this slightly and define $R_L^2 = SSR_1/SSR_2$. The subscript L indicates we are dealing with logistic regression. In ordinary regression SSR can equal $SSTOT$ if all observations are on a line and then $R^2 = 1$. For logistic regression SSR_1 can never be larger than SSR_2. Hence SSR_2 is chosen as the denominator in the definition of R_L^2 (Hosmer and Lemeshow 1989). R_L^2 gives us essentially the same information that the test of fit statistic $SSR_2 - SSR_1$ does. The closer R_L^2 is to 1 the better the logistic model fits the data.

Table 2—Summary information for beetle insecticide example (example 1)[1]

		Number killed		Number surviving		Fraction killed	
Dosage	Number exposed	*O*	*E*	*O*	*E*	*O*	*E*
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1.6907	59	6	3.46	53	55.54	0.1017	0.0586
1.7242	60	13	9.83	47	50.17	.2167	.1639
1.7552	62	18	22.44	44	39.56	.2903	.3620
1.7842	56	28	33.89	28	22.11	.5000	.6052
1.8113	63	52	50.09	11	12.91	.8254	.7951
1.8369	59	53	53.29	6	5.71	.8983	.9032
1.8610	62	61	59.22	1	2.78	.9839	.9552
1.8839	60	60	58.74	0	1.26	1.0000	.9790
	481	291		190			

[1]Columns headed *O* are from observed data; columns headed *E* are from the logistic model (expected).

Example 1.—This example illustrates the use of logistic regression in a dose/response situation. The data come from Stukel (1988) and are given in table 2, columns 1-3, 5, 7. The treatment is CS_2 applied to beetles. Here Y = 0 if a beetle survives and Y = 1 if a beetle dies. The explanatory variable is x = dosage of CS_2, which has only 8 values, and $\pi(x)$ is the probability that an exposed beetle dies as a function of x. Summary statistics and the logistic model follow:

n = # exposed = 481, # killed = 291, # surviving = 190,
$\hat{\pi} = 291/481 = 0.605$

$$SSTOT = -2[291\ell n(291/481) + 190\ell n(190/481)] = 645.44$$

$$\hat{\pi}(x) = 1/(1 + exp(60.7175 - 34.27x))$$

There are eight distinct values $\hat{\pi}(x)$ corresponding to the eight dosages (column 8).

For the logistic model we find

$$SSE_1 = \sum_{i=1}^{481} S_1(y_i, \hat{\pi}_i)$$

$$= -2[6\ell n(.0586) + 53\ell n(0.9414) + \cdots + 60\ell n(0.9790) + 0\ell n(0.0210)]$$

$$= 372.47$$

$$SSR_1 = SSTOT - SSE_1 = 645.44 - 372.47 = 272.97$$

The following ANOVA table summarizes these results. The P-value of 0.000 indicates there is strong evidence to reject Ho: $\beta_1 = 0$.

ANOVA

Source	df	*SS*	*P-value*
Logistic model	1	272.97	0.000*
Residual	479	372.47	
Total	480	645.44	

* from chi-square with 1 df

Model 2 can be found in column (7) of table 2. For model 2 we find

$$SSE_2 = -2[6\ell n(6/59) + 53\ell n(53/59) + \cdots + 60\ell n(60/60) + 0] = 361.24$$

$$SSR_2 = SSTOT - SSE_2 = 284.20$$

Test-of-fit statistic = $SSR_2 - SSR_1 = 284.20 - 272.97 = 11.23$.

The probability of observing a value as large as 11.23 (*P-value*) is found (using a chi-square distribution with df = $k-2 = 6$) to be 0.0815. The *P-value* of 0.0815, although moderately small, does not seem to indicate major discrepancies between model 1 and the data. That is, the logistic model gives a fairly reasonable fit to the data. By this we mean that the $\hat{\pi}$ values calculated from the logistic regression model for the eight dosages of the data set are quite close to the actual observed fraction of the beetles killed. This can be seen by comparing the E and O columns in table 2. In this example, since there are approximately 60 observations at each of the eight dosages, the test of fit above is meaningful. From the information given above we find that R_L^2 = 272.97/284.20 = 0.960. Although we do not interpret R_L^2 quantitatively, it again indicates a pretty good fit of the logistic model.

Additional information is summarized in table 2. The observed column 7 is model 2. The expected column 8 is found by substituting the eight dosage values into the logistic regression model. Plots of these two models appear in

figure 4. Also given in the figure are the number of 0's and 1's at each of the eight dosages.

Model 2 has eight parameters (one for each dosage) and can only be used for these dosages; the logistic model has two parameters and can be used for any dosage in the range of the data. In table 2, the values in column 4 are found by multiplying values in column 8 by the values in column 2. The values in column 6 are found by subtracting values in column 4 from those in column 2. Examining the values in pairs of side-by-side columns headed by O and E gives added insight into how well the model fits the data. Data in columns 3-6 will be used to carry out another test of fit for the logistic model shortly.

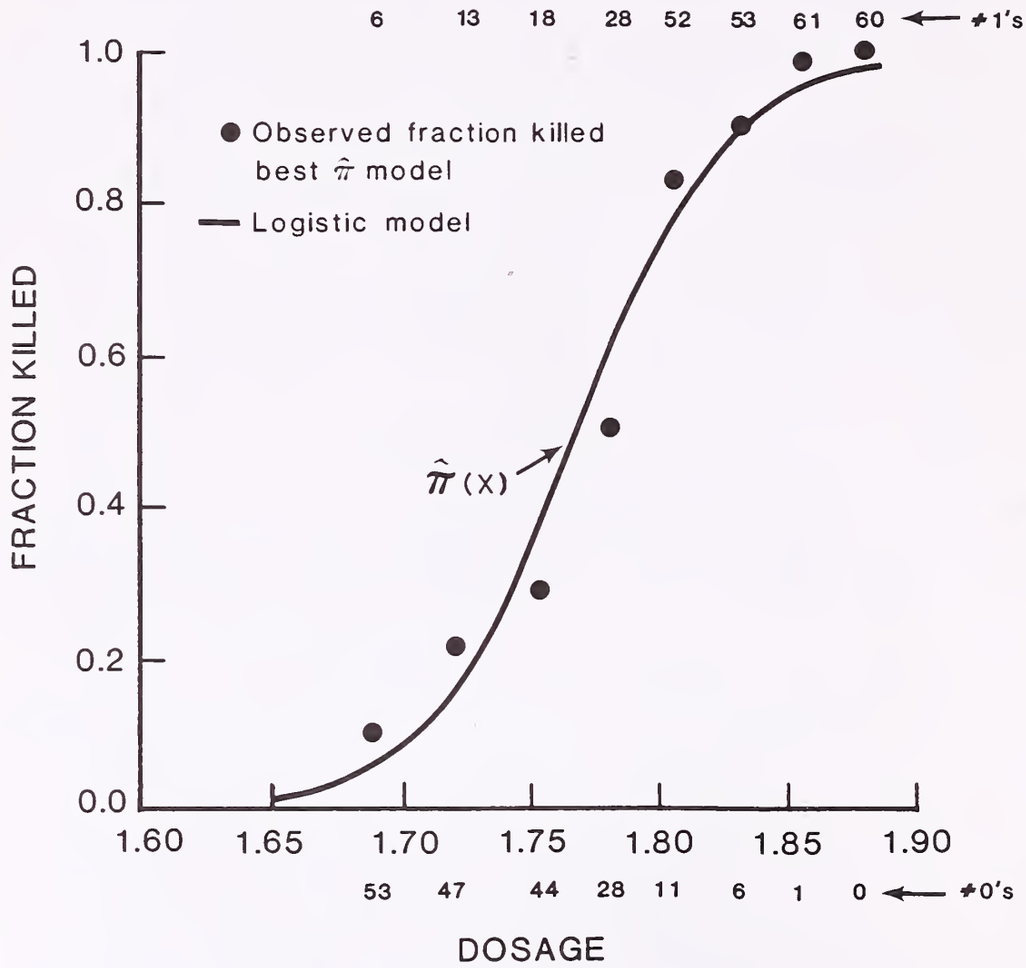


Figure 4—Logistic and best $\hat{\pi}$ models for beetle insecticide example. The number of 0's and 1's at each of the eight dosages are given. The data are summarized in table 2.

Test of Fit Using Grouping Method 1–by Intervals of the Explanatory Variable

When the explanatory variable has many distinct values, some form of grouping is necessary to come up with a reasonable test of fit. The first method we discuss (grouping method 1) groups the data either by distinct values of the explanatory variable if there are not too many distinct values or by disjoint intervals of the explanatory variable. Two other methods will be discussed later (grouping methods 2 and 3).

A chi-square test-of-fit statistic can be calculated using observed and expected numbers of 0's and 1's. It is convenient to think of the data as being summarized in a two-way table where rows correspond to values of Y (0 or 1) and columns are the intervals (or values) for the explanatory variable. Table 3 shows this layout and introduces the notation to be used. O_{ij} is the number of observations in row i and column j, $O_{i\cdot}$ is a row total and $O_{\cdot j}$ is a column total. As an example, the two rows in table 3 correspond to columns 5 and 3, respectively, in table 2. The column headings correspond to the explanatory variable values in column 1.

Table 3—Layout for observed values used in chi-square test of fit (grouping method 1)

Y	1	2	3	...	k	Total
0	O_{01}	O_{02}	O_{03}		O_{0k}	$O_{0\cdot}$
1	O_{11}	O_{12}	O_{13}		O_{1k}	$O_{1\cdot}$
Total	$O_{\cdot 1}$	$O_{\cdot 2}$	$O_{\cdot 3}$		$O_{\cdot k}$	n

Expected values can be computed as follows: Pick a column, say 3. There are $O_{\cdot 3}$ observations in this column, which can be thought of as $O_{\cdot 3}$ Bernoulli trials. The probability that $Y = 1$ in this column can be estimated using the logistic model. Call this probability $\hat{\pi}_3$. Then the expected number of 0's and 1's in column 3 are computed as $E_{03} = (1-\hat{\pi}_3)\cdot O_{\cdot 3}$ and $E_{13} = \hat{\pi}_3 \cdot O_{\cdot 3}$, respectively.

It is these expected and observed values that are used in computing the chi-square statistic. This statistic has an approximate chi-square distribution with $k-3$ degrees of freedom where k is the number of distinct x values or intervals of x values and is computed as follows:

$$\chi^2 = \sum_{i=0}^{1} \sum_{j=1}^{k} [(O_{ij} - E_{ij})^2 / E_{ij}].$$

For example 1 above, we use columns 3 through 6 of table 2 and obtain

$$\chi^2 = (6 - 3.46)^2/3.46 + (13 - 9.83)^2/9.83 + \cdots + (0 - 1.26)^2/1.26 = 10.02.$$

The test statistic in this example has approximately a chi-square distribution with 5 degrees of freedom, which yields a *P-value* of 0.0747. This *P-value* is nearly the same as the 0.0815 for the earlier test of fit. The conclusion is the same, namely, there is some sign of lack of fit with this low *P-value*, but there does not seem to be any major problem when we examine columns 3-6 in table 2. Here again we are comparing observed values (model 2) with expected values (from the logistic regression model 1).

We next offer some examples and further discussion of the tests introduced above. Examples 2 and 3 use data from the Lolo National Forest in Montana. Example 4 compares the 1978 NFDRS with the 1988 NFDRS using data from the Black Creek National Forest in Mississippi for one fire index.

Example 2.—The data come from $n = 2501$ days during the 1970-1985 fire seasons from the Lolo National Forest in Montana. For this example x is an NFDRS index, Energy Release Component (ERC), using fuel model G. We will denote x as ERC(G). The dependent variable Y is 1 if one or more fires are discovered on a day (fire-day) and 0 otherwise (no-fire-day).

The data have been grouped into 12 sets using disjoint intervals of ERC(G). The value of x for all observations in one of these intervals is taken as the midpoint of the interval. The data are summarized in columns 1-4, 6, and 8 of table 4. Some of the results and the logistic model follow.

Table 4—Logistic model using ERC(G) and grouped data (example 2)[1]

ERC(G) range	ERC(G) midpoint		Fire-days		No-fire-days		Fraction fire-days	
		Days	*O*	*E*	*O*	*E*	*O*	*E*
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
0- 0	0.0	11	1	.2	10	10.8	0.091	0.022
1-12	6.5	126	8	5.1	118	120.9	.063	.041
13-17	15.0	119	12	10.5	107	108.5	.101	.088
18-26	22.0	442	71	70.9	371	371.1	.161	.160
27-29	28.0	278	68	70.9	210	207.1	.245	.255
30-32	31.0	306	91	96.1	215	209.9	.297	.314
33-35	34.0	300	96	113.9	204	186.1	.320	.380
36-39	37.5	321	156	148.4	165	172.6	.486	.462
40-45	42.5	379	240	220.9	139	158.1	.633	.583
46-49	47.5	119	78	82.6	41	36.4	.655	.694
50-54	52.0	77	58	59.9	19	17.1	.753	.778
55-59	57.0	23	20	19.6	3	3.4	.870	.851
Total		2501	899		1602			

[1]Columns headed O are from observed data; columns headed E are from the logistic model (expected).

n = # days = 2501, #1's = # fire-days = 899,
#0's = # no-fire-days = 1602

$$SSTOT = -2[899\ell n(899/2501) + 1602\ell n(1602/2501)] = 3266.83$$

Logistic model:
$\hat{\pi}(x) = 1/(1 + exp(3.790 - 0.09705x))$
$SSE_1 = 2805.76, \quad SSR_1 = 461.07, \quad R_L^2 = 0.970$

Model 2:
$SSE_2 = 2791.63, \quad SSR_2 = 475.20$
$SSR_2 - SSR_1 = 475.20 - 461.07 = 14.13$

ANOVA

Source	df	*SS*	*P-value*
Logistic model	1	461.07	0.0000
Residual	2499	2805.76	
Total	2500	3266.83	

Using $SSR_1 = 461.07$ from the ANOVA table which has *P-value* = 0.0000, we have strong evidence to reject H_o:$\beta_1 = 0$. The test-of-fit statistic $SSR_2 - SSR_1 = 14.13$. Using the chi-square distribution with 10 df, the *P-value* is 0.1671. Because this *P*-value is not extremely small, this indicates a good fit for the logistic regression model. Here $R_L^2 = 461.07/475.20 = 0.970$, again indicating a good fit for the logistic regression model.

The chi-square statistic computed using observed and expected fire-days and no-fire-days as given in table 4 is 15.33, df = 9, and *P-value* = 0.0822. Although the *P-value* is small, it is not extremely small, and again we find reasonable agreement between the logistic model and model 2 (the observed data). We emphasize that the logistic model is constructed using all data and can be used for any value of x in the range of the data, whereas model 2 is just the fraction of fire-days at 12 values of x.

Because there were many distinct values of ERC(G), the data were grouped by intervals of ERC(G) to get a reasonable model 2 against which to compare the logistic model. For example 2, the logistic model was also constructed using the grouped data so both models were constructed using the identical data. Generally, one prefers to construct the logistic model using the raw data (ungrouped) because information is lost when the data are grouped as above. Example 3 repeats this example using the raw data to construct the logistic model. Grouping the data to produce a reasonable model 2 against which to compare the logistic model is still necessary. It should be noted that since the raw data logistic model and model 2 based on the grouped data are constructed from slightly different data, it is no longer true that model 2 must be the better model in terms of explained sum of squares. Ordinarily there is little difference between the logistic model based on raw data and that based on grouped data. We recommend using the ungrouped or raw data logistic regression model.

Example 3.—The data and grouping for this example are the same as in example 2. In this example, however, the logistic model is constructed from the raw data while model 2 is still based on the grouped data (see table 5 and fig. 5). Summary results and the logistic model follow:

n = # days = 2501, #1's = #fire-days = 899,
#0's = # no-fire-days = 1602

$SSTOT = 3266.83$

Logistic model:
$\hat{\pi}(x) = 1/(1 + exp(3.8902 - 0.0999x))$
$SSE_1 = 2809.02, \quad SSR_1 = 457.81$

Model 2:
$SSE_2 = 2791.63, \quad SSR_2 = 475.20, \quad R^2_L = 457.81/475.20 = 0.963$
$SSR_2 - SSR_1 = 17.36$, *P-value* $= 0.0668$

ANOVA

Source	df	*SS*	*P-value*
Logistic model (raw data)	1	457.81	0.0000
Residual	2499	2809.02	
Total	2500	3266.83	

The chi-square statistic computed using observed and expected fire-days and no-fire-days is 16.05, df = 9, and *P-value* = 0.0659. From the ANOVA table we have strong evidence to reject H_0:$\beta_1 = 0$ (*P-value* = 0.0000). The *P-value* for the two test-of-fit statistics are smaller than those in example 2, thus showing more indication of lack of fit. We must remember, however, that the logistic model is based on raw data and we are comparing it to the data that has been grouped, thus the fit may not be as good. We prefer the logistic model based on the raw data. We will subsequently introduce better methods of doing tests of fit.

Table 5—Logistic model using ERC(G) and ungrouped data (example 3)[1]

ERC(G) range (1)	ERC(G) midpoint (2)	Days (3)	Fire-days *O* (4)	Fire-days *E* (5)	No-fire-days *O* (6)	No-fire-days *E* (7)	Fraction fire-days *O* (8)	Fraction fire-days *E* (9)
0- 0	0.0	11	1	.2	10	10.8	0.091	0.020
1-12	6.5	126	8	4.7	118	121.3	.063	.038
13-17	15.0	119	12	10.0	107	109.0	.101	.084
18-26	22.0	442	71	68.7	371	373.3	.161	.155
27-29	28.0	278	68	69.8	210	208.2	.245	.251
30-32	31.0	306	91	95.3	215	210.7	.297	.311
33-35	34.0	300	96	113.7	204	186.3	.320	.379
36-39	37.5	321	156	149.0	165	172.0	.486	.464
40-45	42.5	379	240	222.8	139	156.2	.633	.588
46-49	47.5	119	78	83.5	41	35.5	.655	.702
50-54	52.0	77	58	60.6	19	16.4	.753	.787
55-59	57.0	23	20	19.7	3	3.3	.870	.859
Total		2501	899		1602			

[1]Columns headed *O* are from observed data; columns headed *E* are from the logistic model (expected).

Statistical packages for logistic regression use a defacto grouping of the data by the distinct values of the explanatory variable. This grouping gives model 2, and many packages automatically compute a statistic, which in the notation of this paper is $SSR_2 - SSR_1$. As mentioned earlier, if there are many distinct values for the explanatory variable and few observations at each value (possibly only one), this can be a poor model against which to compare the logistic model for a test of fit. With two or more explanatory variables, the problems are usually worse. If both variables have many distinct values, the number of distinct pairs of values and hence the defacto number of groups can be very large. The sum of squares = $SSR_2 - SSR_1$ is routinely reported as a test of fit for many logistic regression programs. It is not based on grouped data and therefore should be used and interpreted with care. It may be useful in situations such as that in example 1 where there are many observations at each of a few distinct values of the explanatory variable, but in general it is useless as a test-of-fit statistic.

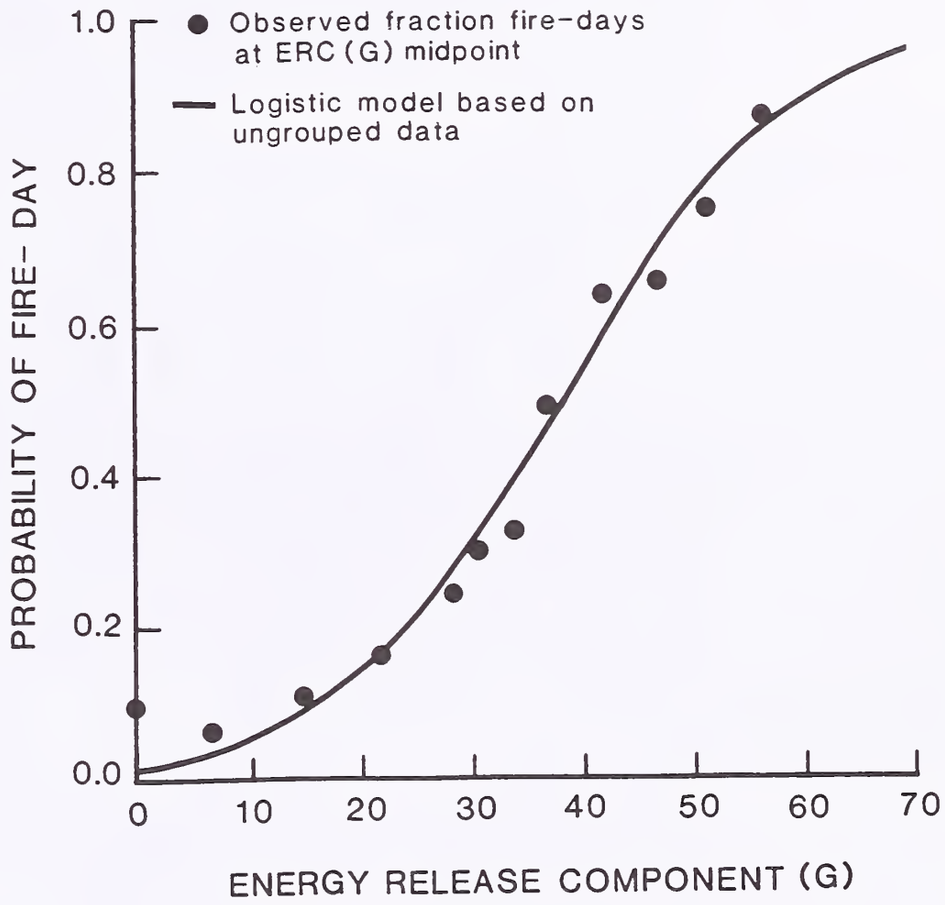


Figure 5--Logistic model based on ungrouped data and observed fraction of fire-days (Lolo National Forest 1970-1985) for twelve ERC intervals. The data are summarized in table 5.

Example 4.—This example illustrates another use for the techniques of this report. The weather and fire data come from Black Creek National Forest in Mississippi. The National Fire Danger Rating System (NFDRS) index to be

used is Burning Index based on fuel model C, which we denote as BI(C). Using the weather data, we generated two sets of BI(C) values, one based on the 1978 NFDRS and the other based on improvements in the NFDRS proposed in 1988. Our goal is to see if the 1988 NFDRS results in detectable improvement in our logistic models. Our approach was to construct and test two logistic models: model A based on the 1978 NFDRS and model B based on the 1988 NFDRS. The two models are then compared to see if the 1988 NFDRS logistic model is "better" than the 1978 NFDRS logistic model in the sense of fitting the data. Summary information and results are now given for each model. The technique is that of example 3.

Model A: Based on 1978 NFDRS models (the data are given in table 6).
n = # days = 3711, # fire-days = 924, # no-fire-days = 2787
$SSTOT = 4165.39$, $SSE_1 = 3661.59$, $SSR_1 = 503.8$

Logistic model:
$\hat{\pi}(x) = 1/(1 + exp(2.1073 - 0.088x))$
$SSR_2 - SSR_1 = 44.38$ df = 8 *P-value* = 0.0000

ANOVA

Source	df	*SS*	*P-value*
Logistic model	1	503.80	0.0000
Residual	3709	3661.59	
Total	3710	4165.39	

The chi-square statistic based on observed and expected fire-days and no-fire-days is 343.21. Using chi-square with df = 7, the *P-value* is 0.0000. A large piece of this statistic is due to the very last cell where observed no-fire-days is 13 and expected no-fire-days is 0.51. Nevertheless, even if this cell is not used, the statistic is still large. The evidence is strong (as measured by the two very small *P-values* for the test of fit statistics) that this logistic model does not fit the data very well. A few very large values of BI(C) have a big influence on the model and results.

Model B: Based on 1988 NFDRS models (the data are given in table 7).
The 1988 NFDRS option of setting 1-hour fuel moisture = 10-hour fuel moisture was used in this example. One immediate result of the 1988 changes was to cut the range of BI(C) values from 0-106 (1978) to 0-44 (1988).

n = # days = 3711, # fire-days = 924, # no-fire-days = 2787

$SSTOT = 4165.39$, $SSE_1 = 3742.88$, $SSR_2 = 422.51$

Logistic model:
$\hat{\pi}(x) = 1/(1 + exp(2.1209 - 0.0981x))$
$SSR_2 - SSR_1 = 3.70$ df = 8 *P-value* = 0.8831

The data and further information for model B are given in table 7.

ANOVA

Source	df	*SS*	*P-value*
Logistic model	1	422.51	0.0000
Residual	3709	3742.88	
Total	3710	4165.39	

For this example, the chi-square statistic based on observed and expected data is 7.51, df = 7, *P-value* = 0.3778. This logistic model fits the data very well according to either test of fit statistic. A careful examination of all information presented for models A and B shows that model B (1988) is far better than model A (1978). This includes examining the observed versus expected values in tables 6 and 7 for the two models, the test statistics and corresponding *P*-values for each model. Figure 6 shows both logistic models as well as the observed fraction of fire-days at the BI(C) midpoints.

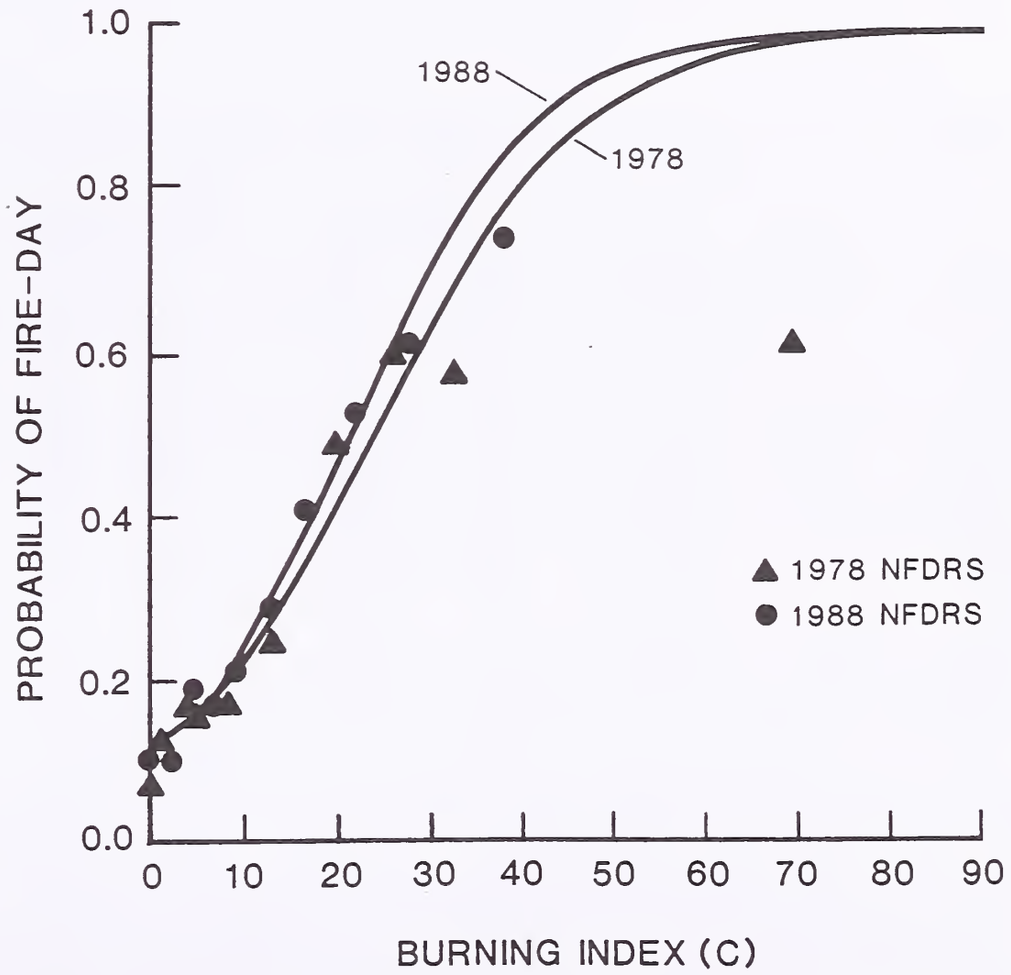


Figure 6—Logistic models based on the 1978 NFDRS and 1988 NFDRS, Black Creek National Forest.

Table 6—Logistic model based on BI(C) and 1978 NFDRS ungrouped data (example 4, model A)[1]

BI(C) range (1)	BI(C) midpoint (2)	Days (3)	Fire-days		No-fire-days		Fraction fire-days	
			O (4)	*E* (5)	*O* (6)	*E* (7)	*O* (8)	*E* (9)
0- 0	0.0	479	34	51.9	445	427.1	0.071	0.108
1- 3	2.0	557	72	70.5	485	486.5	.129	.127
4	4.0	580	101	85.5	479	494.5	.174	.147
5	5.0	353	58	56.1	295	296.9	.164	.159
6-10	8.0	450	76	88.8	374	361.2	.169	.197
11-16	13.5	372	94	106.1	278	265.9	.253	.285
17-24	20.5	553	271	234.9	282	318.1	.490	.425
25-29	27.0	200	120	113.3	80	86.7	.600	.567
30-36	33.0	133	77	91.7	56	41.3	.579	.689
37-106	71.5	34	21	33.5	13	0.5	.618	.985
Total		3711	924		2787			

[1]Columns headed *O* are from observed data; columns headed *E* are from the logistic model (expected).

Table 7—Logistic model based on BI(C) and 1988 NFDRS ungrouped data (example 4, model B)[1]

BI(C) range (1)	BI(C) midpoint (2)	Days (3)	Fire-days		No-fire-days		Fraction fire-days	
			O (4)	*E* (5)	*O* (6)	*E* (7)	*O* (8)	*E* (9)
0	0.0	714	75	76.5	639	637.5	0.105	0.107
1- 3	2.0	385	42	49.0	343	336.0	.109	.127
4- 5	4.5	502	88	78.9	414	423.1	.175	.157
6- 7	6.5	299	51	55.3	248	243.7	.171	.185
8-10	9.0	388	83	87.2	305	300.8	.214	.225
11-14	12.5	512	150	148.6	362	363.4	.293	.290
15-20	17.5	549	227	219.8	322	329.2	.413	.400
21-24	22.5	190	100	99.1	90	90.9	.526	.522
25-32	28.5	145	88	96.1	57	48.9	.607	.663
33-44	38.5	27	20	22.7	7	4.3	.741	.840
Total		3711	924		2787			

[1]Columns headed *O* are from observed data; columns headed *E* are from the logistic model (expected).

LOGISTIC REGRESSION—TWO OR MORE EXPLANATORY VARIABLES

The Logistic Regression Model and ANOVA Table

If there are r explanatory variables $x_1, x_2, \ldots, x_r$, the model with one variable may be extended as follows:

$$E(Y \mid x_1, x_2, \ldots, x_r) = \pi(x_1, x_2, \ldots, x_r)$$
$$= 1/(1+\exp(-\beta_0 - \beta_1 x_1 - \ldots - \beta_r x_r))$$

or, equivalently,

$$\ell n(\pi/(1-\pi)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_r x_r.$$

The data consist of n independent $(r+1)$-tuples, $(x_{11}, x_{21}, \ldots, x_{r1}, y_1), \ldots, (x_{1n}, x_{2n} \ldots, y_n)$. Maximum likelihood estimators for $\beta_0, \beta_1, \ldots, \beta_r$ are denoted $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_r$. The approach used to obtain these maximum likelihood estimators is the same as that used for one variable. All of the expressions and definitions generalize easily. In particular, the definitions for *SSE, SSR,* and *SSTOT* are the same. The only difference is that the logistic regression model is now

$$\hat{\pi}(x_1, x_2, \ldots, x_r) = 1/(1 + exp(-\hat{\beta}_0 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 - \cdots - \hat{\beta}_r x_r))$$

and

$$\hat{\pi}_i = \hat{\pi}(x_{1i}, x_{2i}, \cdots, x_{ri}), \quad i = 1, 2, \cdots, n.$$

A logistic model with s variables in it would generate the following ANOVA table:

ANOVA

Source	df	*SS*
Logistic model	s	SSR
Residual	$n-s-1$	SSE
Total	$n-1$	$SSTOT$

We can use SSR in the ANOVA table to test H_o:$\beta_1 = \beta_2 = \ldots = \beta_s = 0$. The *P-value* is found from a chi-square with $df = s$. We reject H_o when the *P-value* is small. This test is rarely of any use.

With two or more variables in the logistic regression model, the number of distinct s-tuples of explanatory variables is usually large; thus the test of fit based on model 2 and $SSR_2 - SSR_1$ used in earlier examples is not useful here. Different tests of fit must be introduced for use with models having two or more variables. These tests of fit are also usable and quite good when we have only one explanatory variable. Before discussing tests of fit, though, we will discuss another use for the ANOVA table that is analogous to an ordinary regression analysis technique.

Test of Model Improvement by Adding Variables

Suppose s variables are in a logistic regression model. Of interest is the improvement in the model (in terms of reduction in SSE or increase in SSR) if a new explanatory variable is added to the model. We proceed as follows:

1. Model 1: s explanatory variables.
Compute $SSTOT, SSR_1$, and SSE_1 for this model.

2. Model 2: one new explanatory variable added to the s in model 1.
Compute SSR_2 and SSE_2 for this model.

3. The improvement in the model is given by
$SSR_2 - SSR_1 = SSE_1 - SSE_2$.

The ANOVA table is as follows:

ANOVA

Source	df	SS
Logistic (s variables)	s	SSR_1
Improvement (1 extra variable)	1	$SSR_2 - SSR_1$
Residual	$n - s - 2$	SSE_2
Total	$n - 1$	$SSTOT$

The sum of squares in row 2 of this ANOVA table has an approximate chi-square distribution with 1 degree of freedom and it can be used to test whether there is a significant improvement in the model with the added variable. One is not limited to adding just one variable at a time. This method is the basis for the stepwise logistic regression procedure in the BMDP statistical package. This is not a test of fit of the logistic model to the data. Example 5 illustrates these ideas.

Example 5—The data are from the same set (Lolo National Forest) as that used in examples 2 and 3. Two explanatory variables will be used in constructing logistic regression models for the probability of a fire-day. These variables are x_1 = maximum daily temperature and x_2 = 1,000-hour fuel moisture. They are defined as they are used in the NFDRS. Model 1 will include x_1 only, and model 2 will include x_1 and x_2. Summary information for these models follows:

n=# days = 2501, $SSTOT$ = 3266.83, # fire-days = 899,
no-fire-days = 1602

Model 1: x_1 alone
$SSE_1 = 2749.49$, $SSR_1 = 517.34$
$\hat{\pi}(x_1) = 1/(1 + exp(7.3769 - 0.08921x_1))$
Model 2: x_1 and x_2
$SSE_2 = 2630.40$, $SSR_2 = 636.43$
$\hat{\pi}(x_1, x_2) = 1/(1 + exp(0.99144 - 0.62619x_1 + 0.26543x_2))$

Improvement (model 2 over model 1) = $SSR_2 - SSR_1 = 119.09$

ANOVA

Source	df	*SS*	*P-value*
x_1 alone	1	517.34	0.0000*
Improvement when x_2 is added	1	119.09	.0000*
Residual	2498	2630.40	
Total	2500	3266.83	

*from chi-square with 1 df

There is a very significant improvement in terms of sum of squares explained when x_2=1,000-hour fuel moisture is added to a model with x_1=maximum daily temperature in it. Note that this is not a test of fit for the logistic regression model, but a test to see if there is significant improvement when a second variable is added.

Tests of Fit Using Grouping Methods 2 and 3–by intervals of $\hat{\pi}$

As mentioned earlier, a test of fit for a logistic model involves grouping the observations. For one independent variable, the distinct values of this variable give one possible grouping. The problems with this grouping when the variable has many values were discussed earlier. For two independent variables, the analogous grouping would be based on all possible pairs of values for the two variables. The problems encountered for one variable with many values are compounded so this grouping is not viable for use in a test of fit.

With one explanatory (independent) variable, a grouping based on intervals of values for this variable was used. For two independent variables, we could divide the values for each variable into intervals (for example, 12 intervals each). This would then divide the observations up among $12 \times 12 = 144$ groups or cells. This almost always results in many cells with few or no observations in them, as well as small expected values in many cells. This method of grouping is also discarded.

We now discuss a method of grouping that is useful in constructing tests of fit for any number of explanatory variables (Hosmer and Lemeshow 1980, 1989; Lemeshow and others 1988). Once a logistic model is developed, a $\hat{\pi}$ value can be attached to each observation and these $\hat{\pi}$ values can then be used to group the observations. There are two ways this can be done.

Grouping Method 2.—Order the observations by $\hat{\pi}$ values and put the smallest 10 percent in one group, the next smallest 10 percent in a second group, etc., giving 10 groups. Because of ties, each group will usually not contain exactly 10 percent of the observations. Ten groups seems to work well, but exactly 10 groups is not mandatory. Hosmer and Lameshow (1989) recommend splitting ties so that the groups have as close to the same number in each group as possible. This method of grouping is used in the logistic regression procedure in BMDP, but there ties are always kept in the same group.

Grouping Method 3.—Choose 10 intervals, say $[0,1],(.1,.2], \ldots, (.9,1]$, and put all observations having a $\hat{\pi}$ value in the same interval into the same group. Again, 10 groups is not mandatory. Equal length intervals are not necessary either. This method can lead to groups that have few or no observations as well as large differences in the number of observations per group.

Grouping method 2 is preferred and will be illustrated here. The idea of a test of fit is to do the grouping and then count the number of observations in each group to be compared with an expected count computed using the logistic regression model. When there is only one explanatory variable, the results are nearly the same for grouping by $\hat{\pi}$ values or by values of the independent variable.

Once the grouping and tallying of 0's and 1's in each group is completed, the results can be summarized in a table. See table 8 where O_{ij} = number in row i and column j. The expected number of 1's in each group is found by summing the $\hat{\pi}$'s for all observations in the group. The expected number of 0's is then found by subtraction. The rationale for computing the expected number of 1's is: Each observation, Y, is a Bernoulli random variable with $P(Y = 1) \doteq \hat{\pi}$ where $\hat{\pi}$ comes from the logistic model. Thus $E(Y) \doteq \hat{\pi}$. The number of 1's in a group is simply the sum of the Y_i's in the group and hence the expected number of 1's is found by summing the corresponding $\hat{\pi}_i$'s in each group. The expected number of 0's and 1's is denoted by E_{ij} in table 8. In general, each observation in a group could have a different $\hat{\pi}$ and almost certainly all observations in a group will not have the same $\hat{\pi}$.

Table 8—Layout of observed and expected values (example 6)

Y	1	2	3	4	5	6	7	8	9	10
0	O_{01} E_{01}	O_{02} E_{02}	.	.	.	.	.	.	.	$O_{0,10}$ $E_{0,10}$
1	O_{11} E_{11}	O_{12} E_{12}	.	.	.	.	.	.	.	$O_{1,10}$ $E_{1,10}$

The following statistic can be used for a test of fit. It has an approximate chi-square distribution with 8 degrees of freedom and is computed as:

$$X^2 = \sum_{i=0}^{1} \sum_{j=1}^{10} (O_{ij} - E_{ij})^2 / E_{ij}.$$

We return briefly to the one variable case before looking at examples. In this case, we grouped by intervals of the explanatory variable (grouping method 1). To find the expected number of 1's for each interval, it was assumed that all observations in an interval had the same $\hat{\pi}$ value–that corresponding to the midpoint of each interval. This works well if observations are spread out fairly uniformly in each interval, but this is not always the case. The method of finding expected values discussed just above can also be used in the one variable case and this will be illustrated below.

Examples 6, 7, 8, and 9 are based on the same data and variables as example 5. The first three of these examples carry out tests of fit for the model with only one explanatory variable, x_1 = maximum daily temperature, in it. The methods to be used for these examples are grouping by intervals of x_1 and computing expected values using the midpoint of each interval (example 6), grouping by intervals of x_1 and computing expected values by summing $\hat{\pi}$'s in each interval as discussed above (example 7), and grouping by $\hat{\pi}$ values and finding expected values by summing $\hat{\pi}$ values in each group (example 8). Finally, the method of example 8 will be applied to the logistic regression model with two variables, x_1 = maximum daily temperature and x_2 =1,000-hour fuel moisture (example 9). This example illustrates that the test of fit being illustrated here works just as easily for a model with several explanatory variables as it does for a model with only one explanatory variable.

Example 6.—Some summary information can be found in example 5. The logistic regression model contains only x_1 = maximum daily temperature and is $\hat{\pi}(x_1) = 1/(1 + exp(7.3769 - 0.08921x_1))$. Table 9 displays the grouped data and expected values (computed using the midpoint of each interval). From this table, $X^2 = 12.33$, df = 7, *P-value* = 0.0902. This indicates a reasonable fit of the logistic model to the data, although the *P-value* is fairly small.

Table 9—Method 1: grouping by values of maximum daily temperature and computing expected values using midpoint of each interval (example 6)[1]

Maximum daily temperature			Fire-days		No-fire-days		Fraction fire-days	
range	midpoint	Days	*O*	*E*	*O*	*E*	*O*	*E*
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
34-50	42.0	96	7	2.5	89	93.5	0.073	0.026
51-59	55.0	274	23	21.4	251	252.6	.084	.078
60-67	63.5	387	53	59.2	334	327.8	.137	.153
68-74	71.0	409	109	106.6	300	302.4	.267	.261
75-78	76.5	310	116	113.2	194	196.8	.374	.365
79-83	81.0	355	166	164.1	189	190.9	.468	.462
84-87	85.5	276	158	155.2	118	120.8	.573	.562
88-91	89.5	194	116	125.6	78	68.4	.598	.647
91-94	92.5	128	94	90.3	34	37.7	.734	.706
95-102	98.5	72	57	57.9	15	14.1	.792	.804
Total		2501	899		1602			

[1]Columns headed *O* are from observed data; columns headed *E* are from the logistic model (expected).

Example 7.—Everything is the same as in examples 5 and 6 except that here the expected values are computed by summing $\hat{\pi}$ values in each interval. Table 10 summarizes these expected values. From this table, $X^2 = 6.02$, df = 7, *P-value* = 0.5374. This test gives very strong evidence of a good fit

of the logistic model to the data as indicated by the small chi-square value and large *P-value*. The refined method of computing expected values shows a better fit for the model.

Table 10—Grouping values of maximum daily temperatures and computing expected values by summing $\hat{\pi}$ values for observations (example 7)[1]

Maximum daily temperature (1)	Days (2)	Fire-days *O* (3)	Fire-days *E* (4)	No-fire-days *O* (5)	No-fire-days *E* (6)
34-50	96	7	3.9	89	92.1
51-58	274	23	23.0	251	251.0
60-67	387	53	60.4	334	326.6
68-74	409	109	108.0	300	301.0
75-78	310	116	114.4	194	195.6
79-83	355	166	164.4	189	190.6
84-87	276	158	153.7	118	122.3
88-91	194	116	124.5	78	69.5
91-94	128	94	90.7	34	37.3
95-102	72	57	55.9	15	16.1
Total	2501	899		1602	

[1]Columns headed *O* are from observed data; columns headed *E* are from the logistic model (expected).

Example 8.—Everything is the same as in examples 5, 6, and 7 except that grouping is done by $\hat{\pi}$ values, putting approximately 10 percent of the observations in each group (grouping method 2). Expected values are computed by summing $\hat{\pi}$ values in each $\hat{\pi}$ interval. Table 11 summarizes these expected values. From this table $X^2 = 2.70$, df = 8, *P-value* = 0.9518. We find an excellent fit of the logistic model to the data as indicated by the small chi-square value and large *P-value*.

Example 9.—The data are those of example 5. The method of example 8 is applied when the logistic model has two variables. Here there are two explanatory variables, x_1 = maximum daily temperature and x_2 = 1,000-hour fuel moisture. The logistic model is

$$\hat{\pi}(x_1, x_2) = 1/(1 + exp(0.99144 - 0.62619x_1 + 0.26543x_2))$$

The grouping is by $\hat{\pi}$ values (grouping method 2) and expected values are computed by summing $\hat{\pi}$'s in each group. Table 12 gives the expected values. From this table, $X^2 = 5.40$, df = 8, *P-value* = 0.7143. This model also fits the data very well. The effort required to carry out this test of fit is the same regardless of the number of explanatory variables.

Table 11—Grouping by values of $\hat{\pi}$ (grouping method 2) and computing expected values by summing $\hat{\pi}$ values for observations in each interval (example 8)[1]

$\hat{\pi}$	Days	Fire-days		No-fire-days	
		O	E	O	E
(1)	(2)	(3)	(4)	(5)	(6)
[0,0.085]	258	18	15.5	240	242.5
(0.085,0.140]	247	27	28.4	220	218.6
(0.140,0.200]	252	38	43.3	214	208.7
(0.200,0.270]	220	57	51.9	163	168.1
(0.270,0.350]	249	77	76.2	172	172.8
(0.350,0.400]	250	91	94.3	159	155.7
(0.400,0.490]	292	133	132.3	159	159.7
(0.490,0.560]	222	118	117.4	104	104.6
(0.560,0.660]	287	173	176.5	114	110.5
(0.660,0.850]	224	167	163.1	57	60.9
Total	2501	899		1602	

[1]Columns headed O are from observed data; columns headed E are from the logistic model (expected).

Table 12—Logistic model with two variables (maximum daily temperature and 1,000-hour fuel moisture); computing expected values by summing $\hat{\pi}$ values for observations in each interval; using grouping method 2 (example 9)[1]

$\hat{\pi}$ interval	Days	Fire-days		No-fire-days	
		O	E	O	E
(1)	(2)	(3)	(4)	(5)	(6)
[0,0.068]	249	11	10.2	238	238.8
(0.068,0.129]	250	24	24.0	226	226.0
(0.239,0.191]	252	33	40.1	219	211.9
(0.191,0.256]	249	54	55.3	195	193.7
(0.256,0.322]	252	81	73.3	171	178.7
(0.322,0.140]	242	85	88.1	157	153.9
(0.410,0.505]	254	113	114.8	141	139.2
(0.505,0.595]	250	148	137.7	102	112.3
(0.595,0.700]	249	160	159.6	89	89.4
(0.700,0.890]	254	190	195.8	64	58.2
Total	2501	899		1602	

[1]Columns headed O are from observed data; columns headed E are from the logistic model (expected).

Additional Comments on Examples 6, 7 and 8.—In example 6, 8.24 of the total $X^2 = 12.33$ comes from one interval, 34-50, where most of the observations are above the midpoint temperature of 42°. In example 7, some of this problem is solved by the way the expected values are computed. This

problem occurs in the first place because of the small expected values for this interval. In example 8, the method of grouping guarantees a fairly large number of observations in each group, and consequently there are no small expected values to inflate the value of X^2. The method of example 8 is the easiest to implement, and it is automatically carried out in the BMDP statistical package. Tables 9, 10, and 11 are displayed here so one can sense how the grouping methods function and how well the models fit the data. The tables are not necessary to carry out the tests, but are informative.

SUMMARY AND RECOMMENDATIONS FOR LOGISTIC TESTING PROCEDURES

1. Several examples illustrated the use of SSR to test $H_o : \beta_1 = 0$, and example 5 discussed the generalization of this test to the case of several variables. This is a useful test, but it is not a test of fit.

2. Example 5 also illustrated the use of SSR for testing whether the addition of a variable to an existing logistic regression model provides a significant improvement in the model. This is not a test of fit.

3. When there is just one explanatory variable that has a modest number of distinct values and there are several observations at each of these values, then the statistic $SSR_2 - SSR_1$ is useful for carrying out a test of fit. Use of this statistic is illustrated in example 1 and other early examples.

4. For the situation described in (3), a chi-square test statistic for comparing observed and expected values was introduced in example 1 and illustrated in other early examples.

5. When the situation described in (3) does not hold and grouping is necessary, the following recommendations are made:

One explanatory variable
- use either disjoint intervals of the explanatory variable or group by $\hat{\pi}$ values using grouping method 2.
- find expected values by summing $\hat{\pi}$ values in each group.

Two or more explanatory variables
- use grouping method 2.
- find expected values by summing $\hat{\pi}$ values in each group.

6. Most of the procedures discussed in this paper can be carried out using statistical packages such as BMDP, SPSSX, SAS, or GLIM. BMDP has a stepwise logistic regression procedure based on SSR and the ANOVA table discussed in this paper. BMDP also does the Hosmer-Lemeshow grouping method 2 and test of fit. For us BMDP seemed most convenient to use, although no attempt was made to systematically compare the various software packages.

QUADRATIC SCORES

Consider $y_1, y_2, \ldots, y_n$ and corresponding π_i's, $\pi_1, \ldots, \pi_n$. Early in this paper two measures of variation $S_1(y, \pi)$ and $S_2(y, \pi)$ were introduced. At that point we argued that $S_1(y, \pi)$ was the natural one to use, and since then it has been used exclusively. We now briefly discuss matters related to $S_2(y, \pi)$. Recall that $S_2(y, \pi) = (y - \pi)^2$. We could have defined various sums of squares using S_2 just as we did for S_1. For example, if a logistic model $\hat{\pi}(x)$ gives $\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_n$, then SSE would be defined as

$$SSE = \sum_{i=1}^{n} S_2(y_i, \hat{\pi}_i) = \sum_{i=1}^{n} (y_i - \hat{\pi}_i)^2$$

Recall that the y_i's are all either 0 or 1. For this reason SSE, based on either S_1 or S_2, is very sensitive to the location of the data with respect to the logistic curve. If most of the data fall at the lower or upper portions of the curve, most of the y_i's will be near the curve and SSE will be small. If most of the data fall in the central portion of the curve, both 0's and 1's are far from the curve and SSE will be large. This is true even for good-fitting logistic models.

Closely related to the sums of squares defined using $S_2(y, \pi)$ is the idea of a quadratic score for an observation and average quadratic score for a set of observations. The latter has been used to measure goodness of logistic models. Here we will argue that average score is not a useful statistic.

Consider $y_1, y_2, \ldots, y_n$ and corresponding $\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_n$ from a logistic regression model $\hat{\pi}$. For each pair $(y_i, \hat{\pi}_i)$ a quadratic score is defined as

$$\text{score}_i = 1 - 2(y_i - \hat{\pi}_i)^2.$$

Scores have values between -1 and 1 inclusive. When y_i and $\hat{\pi}_i$ are close together, score_i is near 1 and when they are far apart, score_i is near -1. Table 13 displays numerical values.

The average score is defined as follows:

$$\begin{aligned}
\text{Average score} &= \sum_{i=1}^{n} [1 - 2(y_i - \hat{\pi}_i)^2]/n \\
&= \sum_{i=1}^{n} [1 - 2 \cdot S_2(y_i, \hat{\pi}_i)]/n. \\
&= [n - 2 \cdot \sum_{i=1}^{n} S_2(y_i, \hat{\pi}_i)]/n \\
&= 1 - 2 \cdot SSE/n
\end{aligned}$$

where SSE is based on S_2.

Table 13—Numerical values for scores

y	1	0	1	1	0	0	1	1	0
$\hat{\pi}$	1	0	0.90	0.5	0.5	0.90	0.10	0	1
Score=$1-2(y-\hat{\pi})^2$	1	1	0.98	0.5	0.5	−0.62	−0.62	−1	−1

The reader should note that when *SSE* is large, average score is small and when *SSE* is small, average score is large. It is possible to have two different models with the same *SSE* and hence the same average score, but with one *SSE* being 10 percent of *SSTOT* and the other *SSE* being 90 percent of *SSTOT*. As mentioned earlier, *SSE* and, consequently, average score are very dependent on where the data fall with respect to the logistic curve. This is the main reason average score is not useful as a statistic for evaluating logistic models.

These ideas are illustrated with two examples. Both estimate the probability of a fire-day as a function of x = ERC(G). Example 10 is based on fire data from only the Missoula District of the Lolo National Forest. Example 11 includes fires from the entire Lolo National Forest. In both cases, ERC(G) is calculated from Missoula weather records.

Example 10.—This example is based on 15 years of data from the Missoula District of the Lolo National Forest. Summary information follows.

n= # days = 2442, # fire-days = 294, # no-fire-days = 2148,
$\hat{\pi}$=294/2442 = 0.1204.

Logistic model:

$$\hat{\pi}(x) = 1/(1 + exp(4.27604 - 0.07034x)).$$

$$SSTOT = \sum_{i=1}^{2442} S_2(y_i, \hat{\pi})$$

$$= \sum_{i=1}^{2442}(y_i - \hat{\pi})^2$$

$$= 2148(0 - 0.1204)^2 + 294(1 - 0.1204)^2$$

$$= 31.14 + 227.47 = 258.61.$$

$$SSE = \sum_{i=1}^{2442}(y_i - \hat{\pi}(x_i))^2 = 243.37, \quad SSR = 15.24.$$

Average score $= 1 - 2 \cdot SSE/2442 = 0.80$.

The data and expected values are found in table 14. The expected values were found using the $\hat{\pi}$ values corresponding to the midpoint of each

interval of ERC(G). From table 14 we get $X^2 = 10.56$, df = 8, *P-value* = 0.2279. The logistic model fits these data quite well.

Table 14—Missoula Ranger District, Lolo National Forest (example 10)[1]

ERC(G) range (1)	Days (2)	Fire-days		No-fire-days		Fraction fire-days	
		O (3)	*E* (4)	*O* (5)	*E* (6)	*O* (7)	*E* (8)
0-8	78	2	1.4	76	76.6	0.026	0.018
9-16	215	9	7.0	206	208.0	.042	.032
17-24	418	17	23.2	401	394.8	.041	.055
25-32	706	58	66.0	648	640.0	.082	.093
33-36	339	52	46.1	287	292.9	.153	.136
37-40	276	45	47.6	231	228.4	.163	.173
41-44	176	48	38.1	128	137.9	.273	.216
45-48	117	25	31.3	92	85.7	.214	.268
49-52	75	23	24.5	52	50.5	.307	.327
53-56	28	10	10.9	18	17.1	.357	.391
57-60	14	5	6.4	9	7.6	.357	.460
Total	2442	294		2148			

[1]Columns headed *O* are from observed data; columns headed *E* are from the logistic model (expected).

Example 11.—This example is based on approximately 15 years of data from the entire Lolo National Forest. (In example 10 the Missoula District of the Lolo National Forest is the entity.) Both examples 10 and 11 use the Missoula weather station so the sample size remains the same. Summary information follows.

n = # days = 2442, # fire-days = 831, # no-fire-days = 1611
$\hat{\pi}$ = 831/2442=0.3403.

Logistic model: $\hat{\pi}(x) = 1/(1 + exp(3.4877 - 0.0911x))$.

$SSTOT = 548.22$, $SSE = 458.55$, $SSR = 89.67$.

Average score = 0.62.

The data and expected values are found in table 15. The expected values were found using the $\hat{\pi}$ values corresponding to the midpoint of each interval of ERC(G). From table 15 we get $X^2 = 10.70$, df = 8, *P-value* = 0.2193. This logistic model fits the data quite well.

The influence of *SSE* on average score is readily seen. Since $SSE = 243.37$ in example 10 is much smaller than $SSE = 458.55$ in example 11, the average score for example 10 is much larger than that for example 11 (0.80 versus 0.62). Both models, however, fit their respective data equally well.

Table 15—Lolo National Forest (example 1)[1]

ERC(G) range (1)	Days (2)	Fire-days O (3)	Fire-days E (4)	No-fire-days O (5)	No-fire-days E (6)	Fraction fire-days O (7)	Fraction fire-days E (8)
0-8	78	3	3.3	75	74.7	0.039	0.042
9-16	215	15	18.7	200	196.3	.070	.087
17-24	418	69	69.1	349	348.9	.165	.165
25-32	706	186	205.3	520	500.7	.263	.291
33-36	339	149	140.6	190	198.4	.439	.415
37-40	276	144	139.4	132	136.6	.522	.505
41-44	176	108	104.7	68	71.3	.614	.595
45-48	117	75	79.4	42	35.6	.641	.679
49-52	75	50	56.5	25	18.5	.667	.753
53-56	28	21	22.8	7	5.2	.750	.814
57-60	14	11	12.1	3	1.9	.786	.863
Total	2442	831		1611			

[1]Columns headed O are from observed data; columns headed E are from the logistic model (expected).

From a different perspective, the majority of the observations for example 10 are near the lower tail of the logistic model, resulting in a small *SSE* and large average score. The range of $\hat{\pi}$ values in example 11 is much larger than in example 10. Many of the observations fall in the middle portion of the logistic curve, resulting in a large *SSE* and smaller average score.

Examining the data and results summarized above for these two examples shows that both are quite good. Average scores tell us nothing about the fit of the two models and would seem to imply that the model for example 10 is much better than that in example 11. We recommend that average score not be used for a test of fit or for comparing models.

As another illustration of the use of average score and its problems, we draw from Martell and others (1987). In this paper, logistic models were constructed for the probability of a fire-day using 17 years of historical data. Models were constructed for people-caused fires (for several causes and five different seasons). These models were field tested by predicting people-caused fires in 1984. The models were judged on the basis of average score derived from 1984 predictions and observations. Typical average scores were 0.991, 0.831, 0.996, 0.999, and 1.0. At first glance, these scores seem impressive. The year 1984, however, was a very wet year with only six people-caused fires.

Consider the early-summer season and people/recreation-caused fires. In 49 days there was one fire (1 fire-day). The data consist of one 1 and 48 zeroes. Because the year was wet, it is likely that the fire indexes had low values nearly all of the time. This would mean that all of the observations were near the lower tail of the logistic curve where $\hat{\pi}$ is flat and near 0. Because most observations are 0's, *SSE* is small and average score is very large. Thus these average scores do not truly indicate how good the models are or how well they fit the data. They merely reflect where the data fell with respect to the

logistic curves. Again, the conclusion is that average score is not a useful statistic to use in evaluating logistic models.

All of the procedures discussed in this paper are designed to test the models using the data that were used to construct them in the first place. Generally, this gives one lots of observations to work with. Thus the problems of few observations and an atypical year are avoided.

REFERENCES

Andrews, P. L. 1987. The National Fire Danger Rating System as an indicator of fire business. In: Proceedings, 9th conference on fire and forest meteorology; 1987 April 21-24; San Diego, CA. Boston, MA: American Meteorological Society:57-62.

Burgan, R. E. 1988. 1988 revisions to the 1978 National Fire-Danger Rating System. Res. Pap. SE-273. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station. 39 p.

Cox, D. R. 1989. The analysis of binary data. 2nd ed. New York: Methuen and Co., Ltd., 236 p.

Deeming, John E.; Burgan, Robert E.; Cohen, Jack D. 1977. The National Fire Danger Rating System–1978. Gen. Tech. Rep. INT-39. Ogden, UT: U.S. Department of Agriculture, Forest Service, Intermountain Forest and Range Experiment Station. 63 p.

Efron, Bradley. 1978. Regression and ANOVA with zero-one data: measures of residual variation. Journal of the American Statistical Association. 73: 113-121.

Haines, Donald A.; Main, William A.; Frost, John S.; Simard, Albert J. 1983. Fire-danger or rating and wildfire occurrence in the northeastern United States. Forest Science. 29(4): 679-696.

Hartford, Roberta A. 1989. Smoldering combustion limits in peat as influenced by moisture, mineral content, and organic bulk density. In: Proceedings, 10th conference on fire and forest meteorology; 1989 April 17-21; Ottawa, Canada. Ottawa, Canada: Forestry Canada: 282-286.

Hosmer, D. W.; Lemeshow, S. 1980. Goodness of fit tests for the multiple logistic regression model. Communications in Statistics–Theoretical Methods. A9(10): 1043-1069.

Hosmer, D. W.; Lemeshow, S. 1989. Applied logistic regression. New York: John Wiley and Sons. 307 p.

Landwehr, J. M.; Pregibon, D.; Shoemaker, A. C. 1984. Graphical methods for assessing logistic regression models. Journal of the American Statistical Association. 79: 61-83.

Latham, D. J.; Schlieter, J. A. 1989. Ignition probabilities of wildland fuels based on simulated lightning discharges. Res. Pap. INT-411. Ogden, UT: U.S. Department of Agriculture, Forest Service, Intermountain Research Station. 16 p.

Lemeshow, S.; Hosmer, D. W., Jr. 1982. A review of goodness of fit statistics for use in the development of logistic models. American Journal of Epidemiology. 115(1): 92-106.

Lemeshow, S.; Teres, D.; Avrunih, J. S.; Pastides, H. 1988. Predicting the outcome of intensive care unit patients. Journal of the American Statistical Association. 83: 348-356.

Martell, D. L.; Otakel, S.; Stocks, B. J. 1987. A logistic model for predicting daily people-caused fire occurrence in Ontario. Canadian Journal of Forest Research. 17: 394-401.

McCullagh, P.; Nelder, J. A.; 1989. Generalized linear models. New York: Chapman and Hall. 511 p.

Pregibon, D. 1981. Logistic regression diagnostics. The Annals of Statistics. 9(4): 705-724.

Ryan, K. C.; Reinhardt, E. D. 1988. Predicting postfire mortality of seven western conifers. Canadian Journal of Forest Research. 18: 1291-1297.

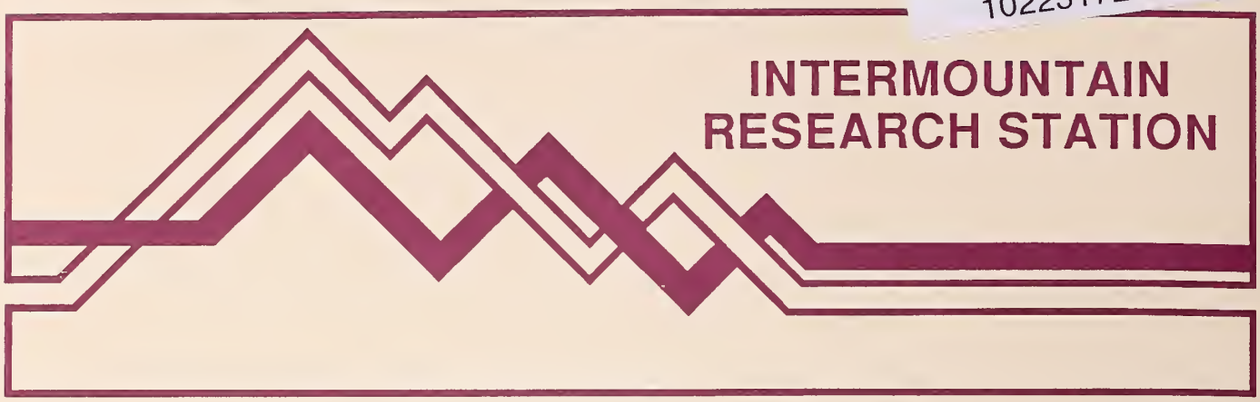
Stukel, Therese A. 1988. Generalized logistic models. Journal of the American Statistical Association. 83: 426-431.

☆U.S. GOVERNMENT PRINTING OFFICE: 1992 - 673-041/41033

Loftsgaarden, Don O.; Andrews, Patricia L. 1992. Constructing and testing logistic regression models for binary data: applications to the National Fire Danger Rating System. Gen. Tech. Rep. INT-286, Ogden, UT: U.S. Department of Agriculture, Forest Service, Intermountain Research Station. 36 p.

Logistic regression was used in examining the relationship between National Fire Danger Rating System (NFDRS) indexes and historical fire occurrence data. Basic techniques of constructing and testing logistic regression models are presented at a modest mathematical level. The emphasis of the paper is on tests of fit for logistic regression models. Specific results of the study are reported elsewhere. The explanatory (independent) variable is an NFDRS index. The response (dependent) variable is fire-day, which has value 1 if one or more fires occur in an area of concern on a certain day and 0 otherwise. This application differs from classic dose/response studies in several ways: there is no control over the "dose," the explanatory variable; indexes may range as high as 200 so it is necessary to group data into categories for tests of fit; and quick and easy methods for doing tests of fit are required for comparing 100 logistic regression models generated from the same set of weather data (20 fuel models x 5 indexes).

KEYWORDS: logistic regression, fire danger rating, fire occurrence



The Intermountain Research Station provides scientific knowledge and technology to improve management, protection, and use of the forests and rangelands of the Intermountain West. Research is designed to meet the needs of National Forest managers, Federal and State agencies, industry, academic institutions, public and private organizations, and individuals. Results of research are made available through publications, symposia, workshops, training sessions, and personal contacts.

The Intermountain Research Station territory includes Montana, Idaho, Utah, Nevada, and western Wyoming. Eighty-five percent of the lands in the Station area, about 231 million acres, are classified as forest or rangeland. They include grasslands, deserts, shrublands, alpine areas, and forests. They provide fiber for forest industries, minerals and fossil fuels for energy and industrial development, water for domestic and industrial consumption, forage for livestock and wildlife, and recreation opportunities for millions of visitors.

Several Station units conduct research in additional western States, or have missions that are national or international in scope.

Station laboratories are located in:

Boise, Idaho

Bozeman, Montana (in cooperation with Montana State University)

Logan, Utah (in cooperation with Utah State University)

Missoula, Montana (in cooperation with the University of Montana)

Moscow, Idaho (in cooperation with the University of Idaho)

Ogden, Utah

Provo, Utah (in cooperation with Brigham Young University)

Reno, Nevada (in cooperation with the University of Nevada)